5.5 A 25W SoC with Dual 2GHz Power™ Cores and Integrated Memory and I/O Subsystems

Zongjian Chen, Priya Ananthanarayanan, Sukalpa Biswas, Brian Campbell, Hao Chen, Shailendra Desai, Shaishav Desai, Dominic Go, Rajat Goel, Vincent von Kaenel, Jason Kassoff, Fabian Klass, Weichun Ku, Tony Li, Jonathon Lin, Khurram Malik, Anup Mehta, Dan Murray, Eric Shiu, Chris Shuler, Sribalan Santhanam, Greg Scott, Junji Sugisawa, Toshinari Takayanagi, Honkai John Tam, Pradeep Trivedi, James Wang, Ricky Wen, John Yong

PA Semi, Santa Clara, CA

The PA6T-1682M SoC targets applications including compute servers, networking, imaging and storage applications [1]. It integrates two 2GHz Power™ architecture cores, a shared 2MB L2, a coherent crossbar interconnect, two 1066MHz DDR2 64b memory channels, a configurable I/O subsystem able to support two 10Gb and four 1Gb Ethernet MAC, eight PCI Express links of configurable width with an aggregate bandwidth of 6GB/s, and hardware acceleration for cryptography, XOR and network functionalities. The functional block diagram is shown in Fig. 5.5.1. The chip die size is 115mm², implemented in a 65nm triple-V_t, dual-oxide 8M CMOS process.

The maximum thermal design power is 25W. To achieve this efficiency, the cores have power-saving modes in addition to various active modes, as shown in Fig. 5.5.2. Each core has an independent supply (V_{DDcpu}), which can be shut down when there is no active workload. This arrangement also enables each core to operate with its own minimum required V_{DD} under the presence of inter-core process and temperature variation. The SRAM arrays have their own V_{DD} supply. The writability of the SRAM cell would otherwise be the limiting factor for the minimum core V_{DD}. The memory and I/O subsystem has its own V_{DD}, which is lower than the technology maximum for additional power savings.

Low-voltage operation exacerbates the impact of PVT variations, so the design flow is enhanced to deal with such variations. Monte Carlo simulations are used to characterize critical circuit elements to optimize device sizing. Post processing of the static timing results compensates for the impact of statistical variations on the margin. Low-V_t cell swapping into critical paths shapes timing histograms and improves speed yield. In this scheme, noncritical devices with a safe design margin are swapped with a longer channel version to reduce leakage power. Longer channel devices are chosen instead of high V_t transistors because of their better voltage scalability and better trade-off of performance versus standby current.

Clock gating is extensively used (23,000 instances) as an intrinsic way to implement logic functions and to save power. In-house tools gather flip-flop toggling statistics at the RTL stage of the design and provide early feedback on the effectiveness of clock gating in each functional unit. As the design progresses towards the physical implementation stage, power is re-estimated with actual parasitic extraction using commercial tools. Good correlation was found between the in-house RTL-level tool and the transistor-level simulation accounting for parasitics, as shown in Fig. 5.5.3.

Two PLLs provide the variable frequencies for the core and fixed frequencies for the I/O and memory subsystem [2]. Debugging functions such as stretch/squeeze/stop of clocks are built into the balanced H-tree clock distribution. The changing core V_{DD} poses challenges for the clocking scheme. The coherent crossbar and functional units it connects to, such as the L2, are clocked at half the core frequency. Due to the adjustable V_{DDcpu}, the clock-tree delay for the core is variable, while the clock-tree delay for the SoC is fixed. Keeping the core synchronous with the coherent crossbar is vital both for reducing L2 and memory access latency and for simplicity of architectural and logic design. The scheme that solves this issue is shown in Fig. 5.5.4. A phase detector detects the phase relationship between the core and the bus clock. The phase is then used to select the circuit path between the core and bus such that the path chosen is synchronous at a fixed architectural delay with maximum possible setup and hold margin. If the path selected needs to change due to a voltage/temperature drift, the bus activity is halted before a switch is made. Hysteresis is built into the path selection hardware to keep the impact on the bus performance to a minimum.

Circuit techniques are used to reduce the latency associated with clock-domain crossing, such as between the SerDes clock and the internal clock domains. The scheme uses an algorithm that periodically detects the phase relationship of the two clocks with enough resolution for the domain transfer to meet setup and hold conditions. In between these detection points, the phase relationship is predicted based on the ratio of the clock periods of the two domains. In the example shown in Fig. 5.5.5, an all-digital phase relationship locker samples (using both rising and falling edges of the writing clock) the divide-by-2 version of the receiving clock, and maintains a lookup table (LUT) index. At each rising edge of the writing clock, the LUT index is initialized when a continuous stream of logic "1" or a stream of "0" of a signature length ends the most recent sampled data. This condition corresponds to a case where the phase relationship between the read and write clock is determined with least error. The LUT index is otherwise incremented. The circuit labeled WptrR Generator passes the write pointer at the edge(s) according to the LUT content indexed. The LUT contents are pre-calculated based on the ratio of read/write clock frequency and the setup/hold margins at the receiving domain. The indexed LUT content at each writing clock edge indicates if this is a safe edge that meets the setup and hold requirement at the receiving clock domain for a pointer transfer.

The 8-way set associative L2 cache has a modular design so that the number of ways and number of banks in each way can be easily reduced for other configurations in the product family. For each bank, 4 pairs of arrays take turns to provide 16B data with its ECC value in 4 consecutive 1ns cycles, forming a 64B cache line. The design takes advantage of the fact that each array will mostly be read once during four cycles, thus the wordline triggering to the sense-amp firing process has a full cycle for a read operation. Write wordline assertion is still a phase operation, as shown in Fig. 5.5.6. Compared to a clock-phase-based design, this approach allocates more time for bitline differential to develop, enabling a longer bitline and hence a more compact design.

The memory controller design allows a system-level power performance trade-off to be made. Energy spent on memory access is the biggest component of energy spent in instruction and data fetch. The memory controller has a mode where ranks are powered down when there are no pending requests. During scheduling, preference is given to the rank that has the most transactions outstanding, thereby maximizing the opportunity to power down the DRAM, achieving significant power savings with very little loss in performance. A proprietary CRC scheme corrects single-bit errors and detects double-bit and burst errors, providing support for reliable operation with detection of DRAM chip failures.

Acknowledgements:
The authors acknowledge the contribution from Tse-yu Yeh, Brian Lilly, Sridhar Subramanian, Mehul Shah, Fernando Aires, Amit Chandra and Matthew Page.

References:
[1] Tse-Yu Yeh, "Low-Power, High-Performance Architecture of the PWRficient Processor Family," *Hot Chips 18*, Aug., 2006.
[2] S. Desai, P. Trived, and V. von Kaenel, "A Dual-Supply 0.2-to-4GHz PLL Clock Multiplier in a 65nm Dual-Oxide CMOS Process," *ISSCC Dig. Tech. Papers*, pp. 308-309, Feb., 2007.

 1-4244-0852-0/07/$25.00 ©2007 IEEE.

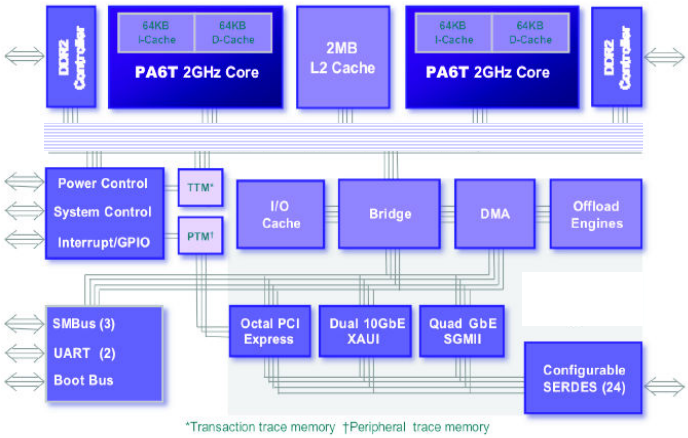

Figure 5.5.1. Chip block diagram.

Power State	Functiodity	Power Consumption	Transition latency
Active	All	25W at 2GHz	NA
Doze	Core clock stopped. Continuous snooping on bus	2.2W	Immediate
NAP	Core clock stopped and voltage lowered. D Cache flushed by hardware. Architecture states retained. SRAM value retained.	1.8W	Entry time 2-16us Wakeup time < 0.5ms
Sleep	Core powered off D-cache flushed by hardware. Software to save necessary architecture states.	1.7W	Entry time 2-16us Wakeup time < 1ms

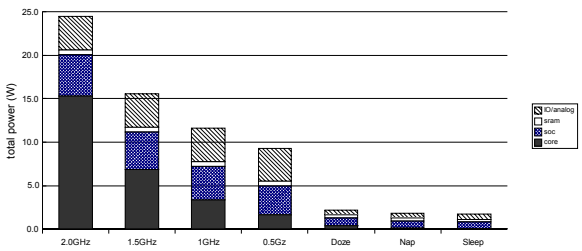

Figure 5.5.2. Power consumption, functionality retained, and exit latency of each power saving state.

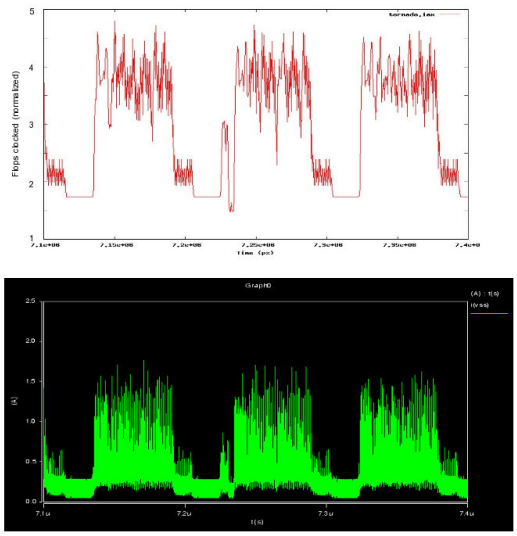

Figure 5.5.3. Power profile comparison between RTL-level (top) analysis and transistor-level analysis (bottom).

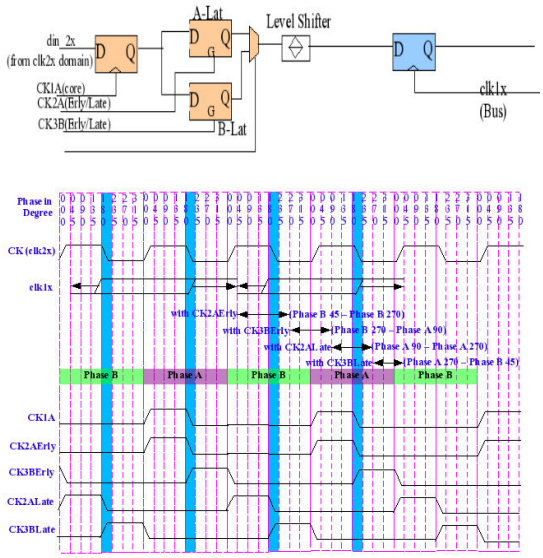

Figure 5.5.4. Core-bus interface synchronization scheme.

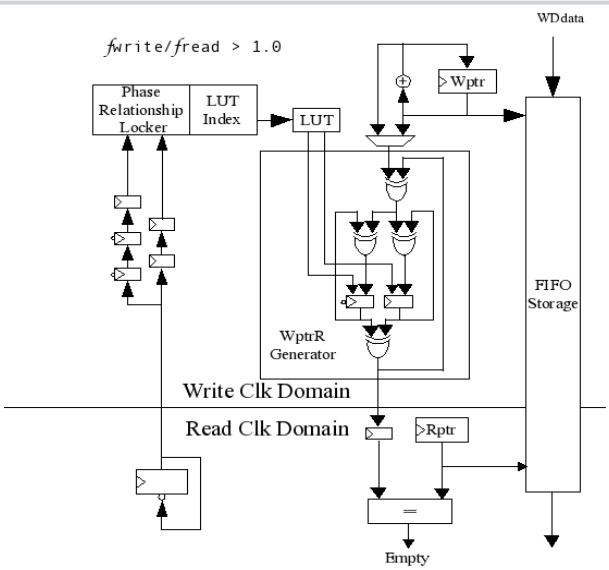

Figure 5.5.5. Low latency synchronizer.

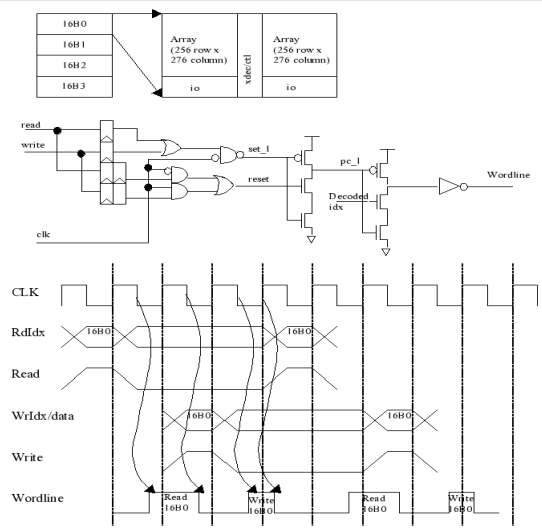

Figure 5.5.6. Cycle-based read wordline operation combined with phase-based write wordline operation in L2 cache optimizes both density and performance.